Bioinformatics for MS Workshop, June 5th,2024, Anaheim

Integrative proteogenomic approach for personalized protein mapping: prospects, challenges, and scalability bottlenecks

Proteogenomics is an area of research at the interface of high-throughput Mass Spectrometry (MS)-based proteomics and Next-Generation Sequencing (NGS)- based genomics. With this approach, customized protein sequence databases generated using genomic and transcriptomic data are utilized to improve the identification of noncanonical and sample-specific proteomes. On the other hand, proteomic data can provide protein/peptide-level evidence of gene expression and help refine gene models.

This workshop explored the advantages and challenges of incorporating various next-generation short— and long-read DNA and RNA sequencing data to generate sample-specific target databases that more accurately reflect the expressed proteome (including, e.g., single-amino acid polymorphisms (SAPs) and alternative splicing events).

We focused on scalability challenges, the need to balance database comprehensiveness with search time constraints, and the inaccurate, false discovery rate (FDR) that comes with searching larger databases. Since the sensitivity and accuracy of proteogenomic detection of neomorphic and noncanonical proteins remain limited by the under-sampling of rare peptides, we also discussed the impact of new MS instrumentation, extensive chromatographic separation, alternative fragmentation mechanisms, and advanced data acquisition strategies for comprehensive mapping of protein diversity.

We introduced an array of computational proteogenomic tools and their utility in the different stages of the proteogenomic pipeline. Furthermore, we reviewed tools that facilitate streamlined data analysis for canonical and noncanonical peptides (e.g., two-pass and cascaded searches). Lastly, we discussed the importance of automated workflows to enable facile sample-specific proteogenomic analyses at scale.

Speakers
Katarzyna Kulej - Memorial Sloan Kettering Cancer Center (MSKCC), New York, United States
Fengchao Yu - University of Michigan, Ann Arbor, Michigan, United States

Panel discussion
Paolo Cifani - Cold Spring Harbor Laboratory (CSHL), New York, United States
Kai Li - University of Michigan, Ann Arbor, Michigan, United States
Helen Mueller – Memorial Sloan Kettering Cancer Center (MSKCC), New York, United States

Summary of workshop themes and discussion:

This workshop was presented in two segments. The first speaker, Katarzyna Kulej, provided a brief overview of the most common proteogenomic methodologies, including (i) the generation of customized protein sequence databases based on various next-generation sequencing techniques; (ii) various aspects of high-throughput and comprehensive proteomics using data-dependent (DDA) and -independent (DIA) acquisition mode. The second speaker, Fengchao Yu, introduced computational tools (i.e., de novo, spectral library search, and sequence database search) and software (i.e., FragPipe) for identifying novel peptides. Lastly, the speakers and invited guests (Paolo Cifani, Kai Li, and Helen Mueller) served as a panel for a roundtable discussion with the audience on the challenges within the

field and on opportunities and strategies for the future. The session was well-attended and featured a wide variety of questions addressing sample preparation, MS acquisition modes, and data analysis.

Next year’s organizers will include Fengchao Yu, who will select the co-coordinator.